Planning for Pascal's Mugging*

Jeffrey Sanford Russell

August 2022

Abstract

In "Pascal's Mugging" (Bostrom 2009), Pascal gives away his wallet for an extremely tiny chance of an extremely large reward. In this continuation of Bostrom's story, Pascal's friend counsels him to take into account the possibility of making mistakes about his true expected utilities, and they consider to what extent this will help Pascal make plans to avoid future muggings.

PASCAL You'll never guess what just happened to me.[1]

ARNAULD Tell me!

PASCAL I just met a kind gentleman in a dark alley who claimed to have magical powers, so that he could grant me any finite amount of happy life at all. And he offered to grant me *1,000 quadrillion days* of happy life, just for giving him my wallet!

ARNAULD Oh dear, I hope you didn't accept that deal. He was almost certainly lying.

PASCAL *Almost* certainly. But what if he told the truth? It's not impossible—just extremely improbable. Indeed, the probability I assigned to him really having such powers and following through was one in ten quadrillion. As you know, I value days of happy life linearly: each additional day is worth 1 util to me. As it happens, my wallet and its contents were also worth 1 util. I told him all this.

*This paper arose from a conversation with John Hawthorne. Thanks also to the Big Decisions working group at USC for helpful discussion.

[1]This is a sequel to the events of Bostrom (2009).

ARNAULD So then he offered to compensate you for your wallet by giving you 1,000 quadrillion days of happy life.

PASCAL Indeed! My expected utility for his offer was 100 utils—probability 10^{-16} times utility 10^{18}—and the cost was only 1 util! What a wonderful opportunity!

ARNAULD *Mon ami,* I am concerned for you. Haven't you heard that there is a notorious gang of muggers on our streets, preying on expected utility maximizers with unbounded utility functions like you? This is how they always proceed: they ask you about your utility function, and the probability you assign to their being honest, and then, whatever you say, they offer enough happy life so that your expected utility will be more than the value of your wallet. You are going to lose your wallet again tomorrow, at this rate!

PASCAL But that's good news! Even more expected utility surplus for me!

ARNAULD It may *seem* so to you at the time. But if you plan ahead now, do you really think you should accept such shady deals?

PASCAL I do. I don't expect my utility function to change between now and tomorrow.

ARNAULD But won't your probabilities change?

PASCAL I suppose so. The mugger I met today had a pale countenance and dark eyes, which he said were the telltale marks of an Operator of the Seventh Dimension. I suppose that when I meet another mugger tomorrow, they won't look precisely the same. The probability I assign to them being honest will be something very small, but it may not be precisely 10^{-16}. It may be 10^{-12} or 10^{-20} or some other small value, depending on the details of the situation.

So I'll make a plan now for how to respond to any such muggers I meet tomorrow. I'll choose my plan based on the expected utility I assign *now* to following the plan *tomorrow*. And, in case it matters, I shall resolutely bind myself to following my chosen plan, even if for some reason I later change my mind.

ARNAULD *Bien.*

PASCAL But this changes nothing. There are various ways a mugger I meet might appear to me—various new pieces of evidence I might gain. Conditional on each possible appearance, I assign some probability to the

mugger's honesty. When I meet the mugger, I shall update my probabilities by conditionalization: my new probabilities will be my *current* conditional probabilities, given that particular way they might appear. I shall tell the mugger my new probability p, and they shall offer me u days of happy life in exchange for my wallet. If $p \cdot u > 1$, then by my *current* lights, accepting the offer has higher expected utility than rejecting it *conditional* on me being in that situation. Thus the expected-utility-maximizing *plan*, by my current lights, is to accept the offer in every such case.

ARNAULD Ah, but here is what worries me. We agree that you ideally *should* update your probabilities by conditionalization. But how confident are you that you will succeed at this? When you meet a mugger in a dark alley, it is hard to be completely confident of exactly how they *appear*, and it is hard to be completely confident of what your conditional probability given their appearance is. Are you sure you won't make a mistake about this?

PASCAL That is a troubling thought.

ARNAULD And this is a situation where a very *small* mistake about your probabilities may matter a great deal. A difference between a probability of 10^{-16} and a probability of 10^{-19} makes the difference between a good deal and a bad one. What's more, it is very difficult to reliably estimate or reason with extremely small probabilities like these.[2]

PASCAL Very well, I will reconsider my plan in light of this. I should consider what kinds of mistakes I am liable to make, and how best to compensate for them.[3]

Let's warm up by considering a very simple model. Let's suppose that there are two ways a mugger I meet might appear to me: *High* evidential probability muggers, who have probability 10^{-16} of being honest conditional on how they appear to me; and *Low* evidential probability muggers, who only have conditional probability 10^{-20} of being honest, given the way they appear. Let's say that I am equally likely to encounter each of these two types of mugger. But I am unreliable at telling them apart from one another. Let's say for the sake of argument that I am

[2]Slovic, Fischhoff, and Lichtenstein (1981) is one classic study among many on cognitive biases that affect estimates of small risks.

[3]Compare the approaches (in other contexts) of Schoenfield (2018); Gallow (2021); Lasonen-Aarnio (forthcoming); Isaacs and Russell (forthcoming).

hopeless at this: regardless of how the mugger appears, I am equally likely to guess High or Low.

ARNAULD Excellent. Then it is straightforward for us to calculate the plan that maximizes expected utility for you. Suppose you are to meet a mugger who offers you a fifty percent return on your "investment".

PASCAL Only fifty percent? The gentleman I met was more generous than that.

ARNAULD Yes, well, let's start there. That means if your estimated probability that the mugger is honest is p, then they offer you $u = 1.5/p$ days of happy life in exchange for your wallet, so $p \cdot u = 1.5$.

In this model your estimate is independent of what kind of mugger you face. So by your current lights, the *conditional expected evidential probability* that the mugger is honest, given your future estimate—whether it is High or Low—is exactly the same as your prior expected evidential probability:

He writes.

$$1/2 \times 10^{-16} + 1/2 \times 10^{-20} = 5.005 \times 10^{-17}$$

If you estimate that the mugger has *High* probability of being honest, the mugger will offer you 1.5×10^{16} days of happy life. So according to your current probabilities, the expected utility of accepting this mugger's offer is just over 0.75 utils. That's a bit less than the one util price they ask you to pay.

PASCAL So, if I estimate that they have High probability of being honest, then even though I will *estimate* that this mugger's offer increases my expected utility by fifty percent, and so I will *think* that accepting the bet maximizes expected utility, I should plan *now* to reject their offer—because I anticipate that my future self will be over-optimistic. Interesting.

But if my estimate is Low?

ARNAULD *Euh* … In this case the mugger offers you 1.5×10^{20} days of happy life, and the expected utility (by your current lights) is … more than $7{,}500$ utils.

PASCAL So even when I make plans that take into account the possibility that I will make mistakes about my expected utilities, I can still take

some opportunities for magnificent expected rewards … when I meet a mugger who strikes me as especially *untrustworthy*? I confess that seems bizarre.

ARNAULD I suppose it does, but here is why. If you *overestimate* the probability that the mugger is honest, then you may judge the mugger's offer to have higher expected utility than you should, by your current lights. But if you *underestimate* the probability that the mugger is honest, then a bet that is *good* by your current lights may seem *bad* to you at the time. If your estimate turns out to be surprisingly *low*, this should make you think it more likely that you have underestimated the true probability, rather than overestimated it. So you should plan to accept the mugger's offer in those cases.

PASCAL I suppose that makes sense. But I have an objection to this simple model. If I know that my estimates are uncorrelated with the true evidential probabilities, then it seems like a big mistake for me to estimate that the probability is Low or High, no matter how things seem to me when I meet the mugger: my estimated probability should rather stand fast at my prior probability ($\approx 5 \times 10^{-17}$) rather than going up or down.

Indeed, it seems that when I face a mugger, I should *then* take into account my *higher-order* evidence about how well my own probability estimates are correlated with the truth of the matter.[4] If I *currently* think that High or Low probability estimate is uncorrelated with whether the mugger is honest, then I should ignore my estimate and just stick to my prior credences. More generally, let's say when I meet the mugger, I *estimate* that the probability on my evidence that they are honest is p. And let's say that my current *conditional* probability that the mugger is honest, given that I *estimate* the probability to be p, is really q. Then I should treat my own estimate as evidence, and update my credence to q, rather than p. Moreover, if I know that this is what I will do, then once again I should plan to accept the mugger's offer in any case.

ARNAULD Well, suppose we grant for the sake of argument that you rationally *ought* to follow this principle about higher-order evidence. Note that this is to give up our previous principle, that you ought to update by conditionalization on your evidence. For we have supposed that you really do *have* the evidence about the mugger's appearance, which raises or lowers the probability of their honesty, even though you may fail to

[4] Compare discussion in Elga (2007); Schoenfield (2018).

correctly update on this evidence.

Still, remember, what we are considering is the possibility that you will update your credences *irrationally*. Before, we considered the possibility that even though you *should* update your credences by conditionalization on your evidence, you might make a mistake and do otherwise. Perhaps instead you should update by conditionalization on your own probability estimate, in the way you suggest. So be it; but isn't this *also* the kind of thing you might make a mistake about? If so, then you are back where you began.

PASCAL I don't think I would make a mistake if things are as in the simple model we have been discussing. If I know I shouldn't change my prior credences at all, this is the sort of thing I think I can carry out without serious error.

ARNAULD Fair enough. In that respect, our model is too simple, then. In your real situation, you are not completely hopeless in assessing your first-order evidence about people's trustworthiness. So you ought to update your credence *some.* And you may well make a mistake about just how much—that is, about what the prior conditional probability of the mugger's honesty is given your future *estimate* of that probability.

PASCAL I take your point. But now let us return to a point from before. According to this simple model, I should plan to devalue my expected value estimate by a factor of 1.9995 in the event that my estimate is High. That protects me from your stingy mugger who offers just 1.5 utils in expectation. But I was made an offer I estimated to be worth 100 expected utils! I should still plan to accept *that* offer.

ARNAULD I did hope to help steel you against more tempting offers. Our simple model represents you as equally likely to make probability estimates that are too low as too high. If your tendency to overestimate probabilities is stronger, then you should plan to resist even more tempting muggings.

PASCAL But if I should really plan to reject even an offer with estimated expected value 100 utils, then my estimates must be *strongly* biased toward overestimates.

ARNAULD A strong bias toward overestimates may not be unrealistic, though. Consider a somewhat less simple model, which may be more reasonable. A *normal* distribution is a natural representation of random errors for

a continuous parameter that can range from $-\infty$ to ∞. We can put probabilities on this scale by considering *log odds*.[5] This is a natural scale for working with probabilities close to zero or one. So a simple but not unreasonable thing to suppose is that your propensity to misjudge evidential probabilities is normally distributed in log odds, with the mean at the *true* posterior log odds. That effectively means you are equally likely to overestimate the true probability by an order of magnitude as you are to underestimate it by an order of magnitude.

PASCAL Those sound like *unbiased* estimates.

ARNAULD That's right—on the log odds scale. But an estimate which is unbiased in log odds is *biased* on the scale of *probabilities* between 0 and 1—so it leads to biased *expected utility* estimates. (See figure 1.) A step *up* by an order of magnitude in probability contributes more to your expected value estimate than a step *down* by an order of magnitude. For example, the average of 10^{-16}, 10^{-15}, and 10^{-14} is greater than the "middle" value of 10^{-15}.

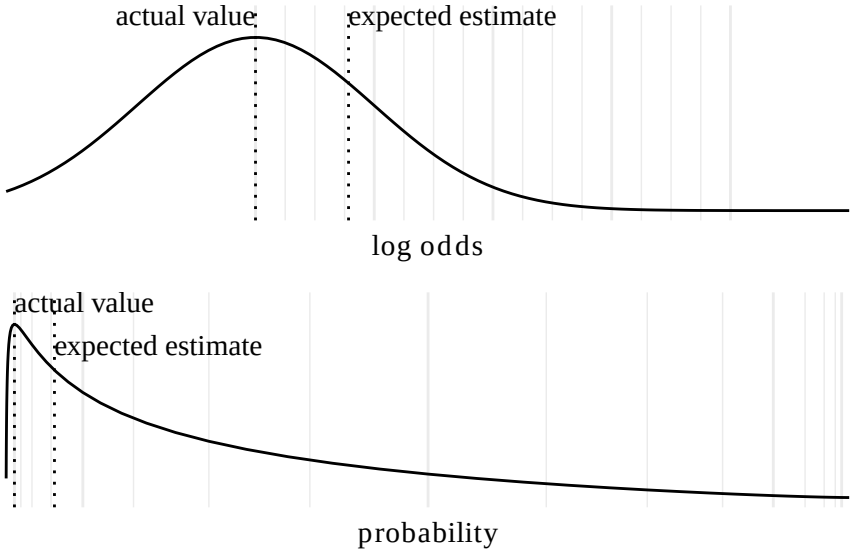

Figure 1: A normal distribution on a log odds scale, and the same distribution on a probability scale from 0 to 1

[5]The log odds of a probability p is $\log\left(p/(1-p)\right)$. For example, a probability of 10^{-16} corresponds to log odds of approximately -16. (We will use base 10 logarithms to keep the correspondence transparent.)

PASCAL I see.

ARNAULD Here's a simple example of how this might work. Suppose that your prior probability for the true evidential probability, given your future evidence, is also normally distributed in log odds. Concretely, suppose the mean of this distribution is -16 and the standard deviation is two orders of magnitude. This means you are about 95% confident that the true evidential probability will be somewhere between 10^{-20} and 10^{-12}.

Suppose that the standard deviation for your error in estimating the evidential probability is a bit wider—three orders of magnitude. In that case, you should plan to downgrade an *estimated* evidential probability of 10^{-12} by a factor of 225. So for that case, at least, you should plan to resist a mugger who offers you 100-fold return.

PASCAL But not if they offer me 1000-fold return.

ARNAULD That's right. If your estimation errors are distributed more widely, then your expected utility discount factor will be even higher. But it will always be finite, so there will be *some* extreme offers you should still plan to accept.

PASCAL And if my estimate is *lower*—say 10^{-20}? Then I should in fact plan to *upgrade* my estimated expected utility of the mugger's offer, right? I should accept an offer from such a mugger even if they promise only 0.1 or even 0.001 expected utils in return, by my estimate.

ARNAULD Yes, I'm afraid so. (See figure 2.)[6]

PASCAL Abstracting from numerical details, I shall resolve to follow this plan. Some muggers will seem more honest to me than others. If I meet a mugger who seem *especially honest*, then I shall decline their offer, generous as it may seem. But if I meet a mugger who seems especially *dishonest*—someone who is suprisingly suspicious-looking, even for a mugger who claims to be a trans-dimensional magician—then I shall *accept* their offer.

ARNAULD I admit that your new plan does not completely reassure me.

Still, we have made some progress. The muggers that roam our streets illustrate two different kinds of problem.

[6]The R source code for these calculations and visualizations can be found at TODO.

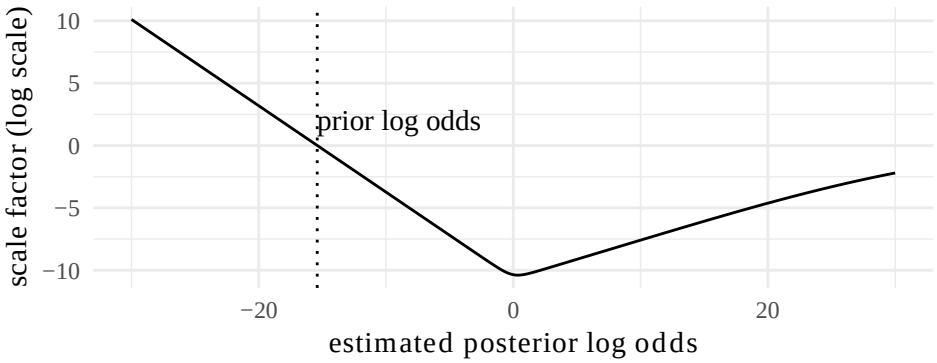

Figure 2: An example of a plan for rescaling estimated expected utilities. The prior over evidential probabilities is normally distributed in log odds with $\mu = -16$ and $\sigma = 2$. Estimated evidential probabilities are normally distributed in log odds with mean at the true evidential probability and $\sigma = 3$. For example, an estimate of 10^{-10} is discounted by 3.7 orders of magnitude, that is, by a factor of about $5{,}000$. An estimate of 10^{-20} is scaled *up* by 3.2 orders of magnitude, or a factor of about $1{,}500$.

> First, there is an *in-principle, theoretical* problem: our standard theory of rational choice advises you to make sacrifices that only have an incredibly tiny probability of resulting in any benefit at all, so long as the utility of that benefit is large enough.[7] This is troubling, and I had hoped that taking into account your propensity to misjudge probabilities would escape this. But that is not how it has turned out thus far. Accounting for errors does make a difference to *what* utilities count as large enough, but it does not change the fact that *some* utilities are large enough.

Pascal Indeed.

Arnauld But these muggers also represent a *quantitative, practical* problem.[8] If you will indulge me, let us imagine future decision-makers of the 21st

[7] This is what Monton (2019, 4) emphasizes:

> No matter how low of a non-zero probability Pascal assigns to the hypothesis that he and the orphans will get the large amount of utility, there is a corresponding utility that he and the orphans could be offered such that Pascal deems the expected utility of the game to be positive, and hence gives the mugger the money.

[8] See Karnofsky (2011) for discussion; compare also Tarsney (2020); Wilkinson (2022); Russell (2021).

century who wish to contribute some of their resources toward doing good for others impartially. I imagine that such people may understand some interventions very well—perhaps (and I am only speculating here) they have special nets that ward off deadly miasmic fevers.

PASCAL This is fanciful, but go on.

ARNAULD Distributing these nets would spare many children from suffering and death. Yet these benevolent beings of the future might instead put their resources toward much more speculative interventions—perhaps an endeavor to build magical machines that, if successful, would benefit far *more* people. And they might be persuaded to do this by assigning a probability to the endeavor's success and calculating expected utilities.

This would not be a case where roving miscreants are promising *whatever* utilities are high enough to exploit a decision-maker. Rather, this is a case of decision-makers simply doing their level best to estimate how much good it might be in their power to do.

PASCAL I see. In such cases, it is very important to recognize that what our decision theory says you *ought to do, given the strength of your evidence* generally comes apart from from what it says you *ought to plan to do upon estimating the strength of your evidence*. Specifically, these two things come apart when you expect your estimates to reflect *biased errors*.

ARNAULD Yes. Moreover, our log odds model suggests a principled reason for expecting biases toward overestimates of very small probabilities. In such cases, decision makers should plan to compensate for their overestimates by effectively "scaling down" their expected utility estimates by some factor.

PASCAL To a point. As we saw, if the estimated probability for the long shot turns out to be surprisingly *low*, then they should scale *up* their expected utility estimate!

ARNAULD So it would seem…

PASCAL In general, whether I should plan to accept or reject a mugger's offer depends not just on *whether* I am inclined to make mistakes about small probabilities, but on what *kinds* and *sizes* of mistakes I am likely to make.

ARNAULD Indeed. Also, besides estimates of evidential probability, we should also consider other kinds of mistakes. How confident are you really about your utilities?

PASCAL Yes, I must consider my plans further.

References

Bostrom, Nick. 2009. "Pascal's Mugging." *Analysis* 69 (3): 443–45.

Elga, Adam. 2007. "Reflection and Disagreement." *Noûs* 41 (3): 478–502.

Gallow, J. Dmitri. 2021. "Updating for Externalists." *Noûs* 55: 487–516.

Isaacs, Yoaav, and Jeffrey Sanford Russell. forthcoming. "Updating Without Evidence." *Noûs*, forthcoming.

Karnofsky, Holden. 2011. "Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased)." The GiveWell Blog. August 18, 2011. https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/.

Lasonen-Aarnio, Maria. Forthcoming. "Perspectives and Good Dispositions." *Philosophy and Phenomenological Research*.

Monton, Bradley. 2019. "How to Avoid Maximizing Expected Utility." *Philosophers' Imprint* 19.

Russell, Jeffrey Sanford. 2021. "On Two Arguments for Fanaticism." *Global Priorities Institute Working Papers Series*, no. 17.

Schoenfield, Miriam. 2018. "An Accuracy Based Approach to Higher Order Evidence." *Philosophy and Phenomenological Research* 96 (3): 690–715.

Slovic, Paul, Baruch Fischhoff, and Sarah Lichtenstein. 1981. "Perceived Risk: Psychological Factors and Social Implications." *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 376 (1764): 17–34.

Tarsney, Christian. 2020. "Exceeding Expectations: Stochastic Dominance as a General Decision Theory." *Global Priorities Institute Working Papers Series*, no. 3.

Wilkinson, Hayden. 2022. "In Defense of Fanaticism." *Ethics* 132 (2): 445–77.